Highly matching Coronavirus-like short sequences can be retrieved from environmental metagenomes

Maximilian Mora
Technische Universitat Graz
Dilfuza Egamberdieva
National University of Uzbekistan
Robert Krause
Medizinische Universitat Graz
Jose Luis Martinez
Centro Nacional de Biotecnologia
Tomislav Cernava (tomislav.cernava@tugraz.at)
Technische Universitat Graz https://orcid.org/0000-0001-7772-4080
Gabriele Berg (gabriele.berg@tugraz.at)
Technische Universitat Graz

Short report

Keywords: COVID-19, false-positive virus detection tests, software, virome studies, DNA sequences, metagenomic data

DOI: https://doi.org/10.21203/rs.3.rs-44155/v1

Abstract

Background

The ongoing COVID-19 pandemic has not only globally caused a high number of causalities, but is also an unprecedented challenge for scientists. False-positive virus detection tests not only aggravate the situation in the healthcare sector, but also provide ground for conspiracy theorists. Previous studies have highlighted the importance of software choice and data interpretation in virome studies. We aimed to further expand theoretical and practical knowledge in virome studies by focusing on short, virus-like DNA sequences in metagenomic data.

Results

Exemplarily, analyses of metagenomic data from the dried-out Aral Sea basin in Uzbekistan showed that Coronavirus-like sequences have existed in environmental samples before the current COVID-19 crisis. In the analyzed environmental datasets, diverse Betacoronavirus-like sequences were detected, which also included SARS-CoV-2 matches. Moreover, the data sets harbored regions that show complementarity to PCR primers that are currently in use for virus detection and origin tracing.

Conclusions

Our study confirms the importance of parameter selection, especially in terms of read length, for reliable virome profiling. Natural environments are an important source of Coronavirus-like nucleotide sequences that should be taken into account when virome datasets are analyzed and interpreted. Moreover, the identified test primer binding sites in metagenomic data might play an important role for the development of reliable test systems and origin tracing.

Background

The worldwide pandemic of COVID-19 (CoronaVirus-Disease-2019) due to SARS-CoV-2 started in 2019 and has since affected several million people and caused hundreds of thousand deaths (WHO). Both numbers were still rising at the time point of concluding this study. It is commonly accepted that SARS-CoV-2 (also known as hCoV19 and 2019-nCoV), the pathogen behind COVID-19, is a zoonotic virus which originated from bats (1, 2) and underwent most likely at least one more recombination event in another intermediate host, before being able to infect humans in large scale (3). A similar mechanism has been already described for two other severe human pathogenic Coronavirus outbreaks, SARS-CoV and MERS-CoV, for which civets (4) and dromedary camels (5) were identified as intermediate hosts respectively. Regarding SARS-CoV-2, the discussion about possible intermediate hosts is still ongoing and so far snakes (3) and pangolins (6) have been suggested. The occurrence of a specific mutation in the receptor binding domain of the spike protein is considered crucial, which consequently allows the virus to dock on human cells as has been already excellently reviewed elsewhere (7). In addition to the unclear origin of SARS-CoV-2, the health sector had been drastically challenged by a scarcity of reliable detection

methods. Indeed, the genome of SARS-CoV-2 is very similar to those of other Coronaviruses, particularly to SARS-CoV. Consequently, false-positive detections of Coronaviruses may complicate epidemiological studies. This is particularly relevant for natural ecosystems, containing complex microbiomes, where viruses can be abundant and diverse (8). Foregoing approaches have demonstrated the importance of assembly strategies for short reads to facilitate reliable virome analyses (9). In the context of the current pandemic, detection of Coronavirus-like sequences, not necessarily belonging to SARS-CoV-2, may compromise the study of the natural distribution of this virus, a feature with relevance for understanding the viral dissemination from a One-Health approach. In this short report we aim to raise awareness of the environmental virome and coronavirus-like sequences. Current data supports the hypothesis that the environmental virome as well as short virus-like sequences are the main reasons for various uncertainties.

Methods

We studied the rhizosphere microbiome of common seablite *Suaeda acuminata* (C.A.Mey.) Moq. to understand the ecology of this first pioneer plant of extreme environments. Two composite samples (A53, A102) analyzed here are part of a larger collection taken in August 2019 near the Large Aral Sea's west shore line in Uzbekistan. DNA was extracted by an established rhizosphere DNA extraction protocol, as described previously (10) and metagenomic shotgun sequencing was conducted by the commercial provider Genewiz (Leipzig, Germany; Illumina HiSeq2500; 2 × 150 bp). Sequencing adapters were removed and reads were trimmed and filtered with trimmomatic (11). The viral taxonomy of high-quality reads was assessed via Kaiju in MEM (maximum exact match) mode (12; viruses-only from the NCBI RefSeq database, compiled on 22.05.2020). Additionally, bowtie 2 (ref. 13, unpaired mode, default parameters, allowing one mismatch) was used to align a list of primers and probes designed for SARS-CoV-2 detection by different internationally renowned institutes and published on the website of WHO (14) against our metagenomic datasets. Sequences containing ambiguous characters were aligned separately for each possible nucleotide exchange. Subsequently, we performed a megablast search of all complete readpairs which were identified by one of the two approaches against various NCBI nucleotide databases: Betacoronavirus GenBank, updated 2020/05/22; Coronaviridae SRA Contigs, updated 2020/05/22; Nucleotide collection (nt), updated 2020/05/21 (ref.15 + 16).

Results And Discussion

Only a small fraction of the datasets could be assigned to viruses with the Kaiju MEM approach. In sample A53 the viral fraction accounted for 0.85% (479,108 of 56,395,621 reads) and in sample A102 it accounted for 0.95% (475,389 of 50,057,128 reads). Of these, 0.01%-0.02% were assigned to *Coronaviridae* in both datasets and a few of them were even classified as SARS-CoV-2 (4 reads in A102 and 3 reads in A53, see Fig. 1). These were short (11 amino acids) exact matches of orf1ab polyprotein (YP_009724389.1), nucleocapsid phosphoprotein (YP_009724397.2), and a membrane glycoprotein (YP_009724393.1). The hits were subjected to a megablast search against NCBI databases and did not

return any significant hits besides the nucleocapsid phosphoprotein read which was together with its paired read also found in the genome sequence of the archaeum *Haloterrigena turkmenica* DSM 5511 (CP001860.1; Nucleotide collection (nt)). Out of the 55 primer and probe sequences published by WHO (14; Supplementary table 1), eleven could be detected at least once in at least one of the two datasets (Table 1). As in most cases only one primer of a primer-pair was detected, we would not expect these hits to lead to positive test results. The only case where forward and reverse primers of the same primer-pair were detected were the primers nCoV_IP2-12669Fw and nCoV_IP2-12759Rv from Institute Pasteur, Paris, France. These were found in both datasets, but the associated qPCR Probe which would also be needed for a clear positive test could not be detected. Generally, all detected sequences were detected at a very low abundance (< 40 reads) with the exception of the qPCR Probe WH-NIC_N-P of the National Institute of Health, Thailand, which was detected in one order of magnitude higher (Supplementary Table 1). Megablast search of the SARS-CoV-2 test-sequence reads found only one unpaired read, which was hit by the 2019-nCoV_N2-R primer (US CDC, USA), that was also identified partly (43/150nt) in a bat metagenome virus discovery project (SRP224019; Coronaviridae SRA Contigs). Additionally, 33.5% (A53) and 42.8% (A102) of SARS-CoV-2 test-sequence reads had reported hits in the nucleotide collection (nt) database while the remaining reads were unknown. Because of the low proportion of viral sequences in the datasets, which is not surprising for metagenomic datasets, it was not possible to assemble longer sequences for more reliable database alignments as previously recommended (9). Nevertheless, we also examined the flanking regions on the short reads identified as SARS-CoV-2 by the Kaiju-MEM approach hit by the SARS-CoV-2 test-sequences and found that they could not be aligned further to SARS-CoV-2 genomes available at GISAID (17).

Table 1
List of SARS-CoV-2 detection primer sequences which can be aligned with at least one of the datasets (A53 and A102). Only positive or negative alignments are shown, for more details see Supplementary Table 1. Two exact WH-NIC N-P alignments were detected, all other alignments were reported with one mismatch.

Primer name	Institute	A53	A102
ORF1ab_R	China CDC, China	+	-
nCoV_IP2-12669Fw	Institut Pasteur, Paris, France	+	+
nCoV_IP2-12759Rv	Institut Pasteur, Paris, France	+	+
2019-nCoV_N1-F	US CDC, USA	+	+
2019-nCoV_N2-R	US CDC, USA	+	+
NIID_WH-1_F501	National Institute of Infectious Diseases, Japan	+	-
NIID_WH-1_F509	National Institute of Infectious Diseases, Japan	+	+
NIID_WH-1_Seq_R24865	National Institute of Infectious Diseases, Japan	+	-
HKU-ORF1b-nsp14F	HKU, Hong Kong SAR	+	-
HKU-NP	HKU, Hong Kong SAR	-	+
WH-NIC N-P	National Institute of Health, Thailand	+	+

Consequently, we argue that, although primers for detecting SARS-CoV-2 and other short sequence matches were found, their finding does not support that the virus was actually present in these samples. We rather detected naturally occurring relatives or Coronavirus-like sequences of which most are of yet unknown origin. The latter is more plausible as we implemented a DNA-based approach. Nevertheless, the fact that Coronavirus-like sequences and even short exact matches to SARS-CoV-2 were detected, hints at the presence of yet unknown environmental viruses or related sequences in this exemplarily analyzed plant-associated habitat. Considering that we also found SARS-CoV-2 test-primer and -probe matches, the existence of such sequences provides a possible explanation for the occurrence of false positive tests as recently reported by news outlets (18, 19) that were welcomed by conspiracy theorists.

Conclusions

Based on our findings we hypothesize in the ecological context that the plant rhizosphere and other natural environments host certain Coronaviruses and/or related nucleotide sequences with a yet unknown role in these ecosystems that remain to be explored in the future. Nevertheless, we still want to stress that the analysis of short sequences from metagenomic data is valuable to identify environmental sequences which might interfere with reliable SARS-CoV-2 detection as most of these test primers and probes have so far only been validated against other known Betacoronaviruses and other viruses that cause respiratory diseases (e.g. 20, 21). As SARS-CoV-2 has been detected already multiple times in

wastewater samples, it has been proposed that wastewater surveillance might be a good strategy for early detection of SARS-CoV-2 outbreaks (22). Our results indicate that profiling of non-assembled virome datasets by standardized microbiome analysis workflows or simple real-time PCR tests are not reliable enough to avoid false positives in complex microbiomes such as those present in soil or wastewater. Specific studies of complex environmental microbiomes will be required, preferably based on RNA-centric approaches with representative sequence lengths, to elucidate the distribution of Coronaviruses and to distinguish them from Coronavirus-like sequences. We are confident that microbiome research can contribute to understand and predict further outbreaks and to implement required actions to avoid them, because viruses are part of the microbiome and their indigenous stability and plasticity is fundamental.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

Metagenomic datasets A53 and A102 are deposited at European Nucleotide Archive (accession numbers: ERR4194679; ERR4194680).

Competing interests

The authors declare that they have no competing interests.

Funding

This work has not received external funding.

Authors' contributions

TC and GB conceived the idea for the study; DE organized and coordinated the sampling; MM performed laboratory experiments and analyzed data; MM, TC, and GB interpreted the data; MM, TC, and GB wrote and edited the manuscript with specific inputs from DE, RK and JLM. All authors read and approved the final manuscript.

Acknowledgements

We appreciate the help of Monika Schneider-Trampitsch, Barbara Fetz, and Julia Kranyecz during wet lab preparations of the samples and thank Dr. Wisnu A. Wicaksono for critically reading the manuscript.

References

1) Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. The species Severe acute respiratory syndrome-related coronavirus: classifying 2019-nCoV and naming it SARS-CoV-2. Nat Microbiol. 2020; 5(4): 536–544.

2) Zhou P, Yang XL, Wang XG, Hu B, Zhang L, Zhang W *et al.* A pneumonia outbreak associated with a new coronavirus of probable bat origin. Nature. 2020; 579(7798): 270-273.

3) Ji W, Wang W, Zhao X, Zai J, Li X. Cross-species transmission of the newly identified coronavirus 2019-nCoV. J Med Virol. 2020; 92(4): 433-440.

4) Wang LF, Eaton BT. Bats, civets and the emergence of SARS. In: Childs JE, Mackenzie JS, Richt JA (Eds.) Wildlife and emerging zoonotic diseases: the biology, circumstances and consequences of cross-species transmission. 1st edn. (Springer, Berlin, Heidelberg, 2007) pp 325-344

5) Azhar EI, El-Kafrawy SA, Farraj SA, Hassan AM, Al-Saeed MS, Hashem AM *et al.* Evidence for camel-to-human transmission of MERS coronavirus. N Engl J Med. 2014; 370(26): 2499-2505.

6) Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol. 2020; 30(8): 1346-1351

7) Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020; 26(4): 450-452.

8) Handley SA, Virgin HW. Drowning in Viruses. Cell. 2019; 177(5):1084-1085.

9) Sutton TD, Clooney AG, Ryan FJ, Ross RP, Hill C.; Choice of assembly software has a critical impact on virome characterisation. Microbiome. 2019; 7(1):12.

10) Köberl M, Müller H, Ramadan EM, Berg G. Desert farming benefits from microbial potential in arid soils and promotes diversity and plant health. PLoS One. 2011; 6(9): e24452.

11) Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014; 30(15): 2114-2120.

12) Menzel P, Ng KL, Krogh A. Fast and sensitive taxonomic classification for metagenomics with Kaiju. Nat Commun. 2016; 7: 11257.

13) Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012; 9(4): 357-359.

14) World Health Organization. Molecular assays to diagnose COVID-19: Summary table of available protocols [Internet]. Available from: https://www.who.int/who-documents-detail/molecular-assays-to-diagnose-covid-19-summary-table-of-available-protocols [accessed 2020 May 11]

15) Zhang Z, Schwartz S, Wagner L, Miller W. A greedy algorithm for aligning DNA sequences. J Comput Biol. 2000; 7:203-214.

16)Sayers EW, Beck J, Brister JR, Bolton EE, Canese K, Comeau DC *et al*. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2020; 48(D1): D9–D16

17) Shu Y, McCauley J. GISAID: Global initiative on sharing all influenza data–from vision to reality. Eurosurveillance. 2017; 22(13): 30494

18) Elliott JK. Global News. Tanzanian president blames lab after goat, papaya 'test positive' for coronavirus [Internet]. Available from: https://globalnews.ca/news/6910821/coronavirus-papaya-goat-tanzania/ [accessed 2020 May 12]

19) Ng K. Independent. Tanzania coronavirus kits raise suspicion after goat and pawpaw test positive [Internet]. Available from: https://www.independent.co.uk/news/world/africa/coronavirus-tanzania-test-kits-suspicion-goat-pawpaw-positive-a9501291.html [accessed 2020 May 12]

20) Corman VM, Landt O, Kaiser M, Molenkamp R, Meijer A, Chu DKW *et al.* Detection of 2019 novel coronavirus (2019-nCoV) by real-time RT-PCR. Eurosurveillance 2020. 25(3): 2000045.

21) Vogels CB, Brito AF, Wyllie AL, Fauver JR, Ott IM, Kalinich CC *et al*. Analytical sensitivity and efficiency comparisons of SARS-COV-2 qRT-PCR assays. medRxiv preprint. doi: https://doi.org/10.1101/2020.03.30.20048108

22) Venugopal A, Ganesan H, Raja SS, Govindasamy V, Arunachalam M, Narayanasamy A, *et al*. Novel Wastewater Surveillance Strategy for Early Detection of COVID–19 Hotspots. Curr Opin Environ Sci Health. 2020; 17:8-13

23) Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC bioinformatics. 2011; 12: 385.

Supplementary Data

Supplementary File 1. Interactive Krona Chart of Figure 1A that includes all reads assigned to viral taxons.

Supplementary File 2. Interactive Krona Chart of Figure 1B that includes all reads assigned to viral taxons.

Supplementary Table 1. More detailed version of Table 1. Listing all names, sequences, originating institutes and hitcounts of primer/probe sequences which could and could not be aligned to our datasets. Two exact WH-NIC N-P alignments were detected; all other alignments were reported with one mismatch.

Figures

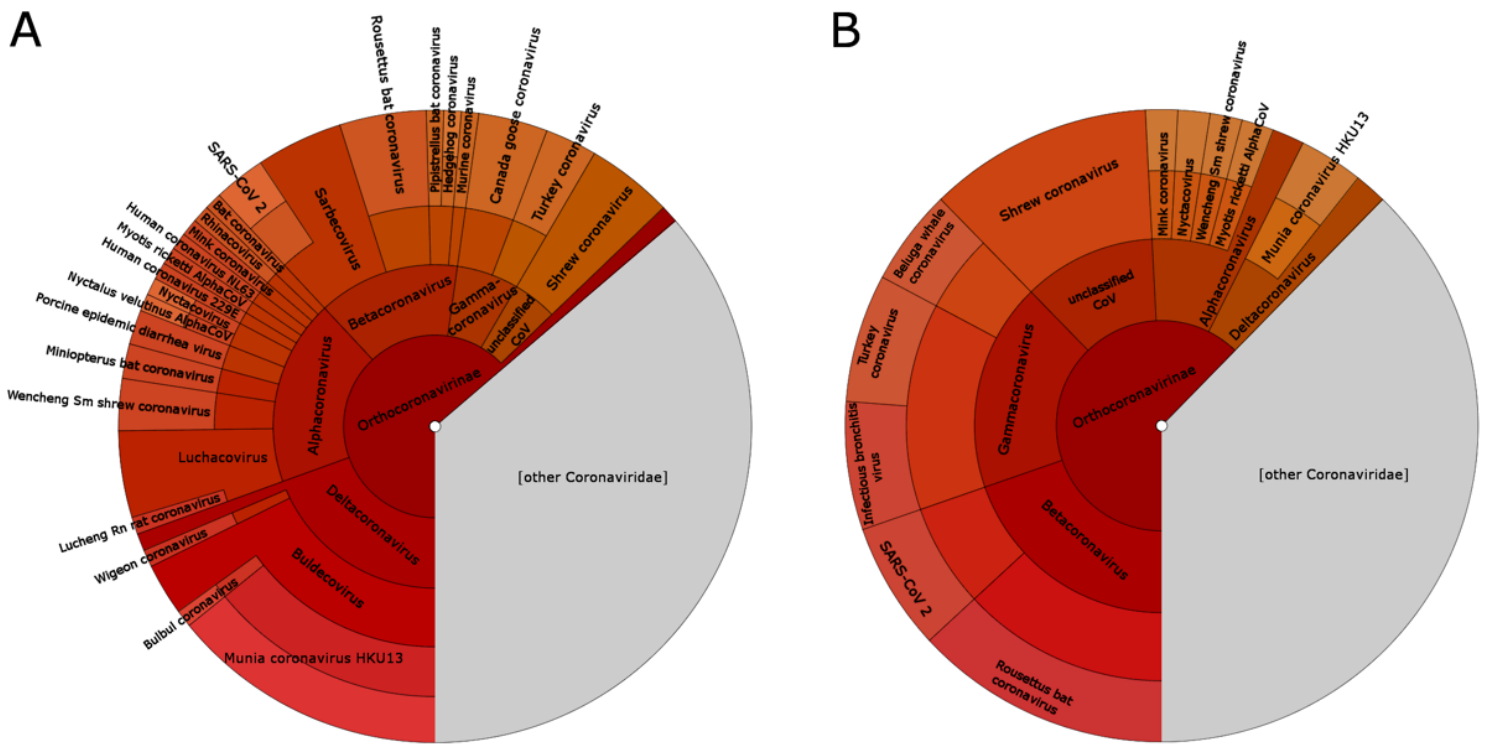

Figure 1

Krona charts with all short-read-based hits for Coronaviridae in plant rhizosphere samples retrieved from the dried-out Aral Sea basin. A: Coronavirus-centric metagenome profile of sample A53. B: Coronavirus-centric metagenome profile of A102. Coronaviridae account for 0.02% (A) and 0.01% (B) of the detected virus sequences. The collapsed snapshots were created with the visualization tool Krona (23) and modified for a better overview. Interactive Krona Charts are available in the supplementary material (Supplementary file 1&2).

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementaryfile2A102krona.html.html
- Supplementaryfile1A53krona.html
- SupplementaryTable1.pdf